

Description of the Dataset for Air Quality Forecast

Introduction

The dataset is comprised of six parts of data that were collected in ([Urban Computing Team](#), Microsoft Research) [Urban Air](#) project over a period of one year (from 2014/05/01 to 2015/04/30), named city data, district data, air quality station data, air quality data, meteorological data and weather forecast data, respectively. This dataset covers 4 major Chinese cities (Beijing, Tianjin, Guangzhou and Shenzhen) and 39 adjacent cities within 300 kilometers to them. Each city is associated with a geo-location denoted by (latitude, longitude), containing a set of districts. In total, there are 2,891,393 air quality records, 1,898,453 (real-time) meteorology records, and 910,576 weather forecast records. Air quality is recorded at 437 air quality stations every hour. The real-time meteorological data are collected at a district (or city) level every hour. Weather forecast has a district (or city)-level record of two coming days, with a temporal granularity of 3 hour, or 6 hour, or 12 hour. Details about these datasets are presented in Table 1.

Table 1. Some Details of Datasets

Datasets		Beijing	Tianjin	Guangzhou	Shenzhen
Time span		2014/5/1-2015/4/30	2014/5/1-2015/4/30	2014/5/1-2015/4/30	2014/5/1-2015/4/30
Nearby cities		14	17	19	19
AQI	In-city stations	36	27	42	11
	In-city instances	278,023	189,604	281,436	88,139
	Ave. PM2.5	106.4	104.3	59.5	44.9
	Neighbor Sta.	233	267	145	148
	#. of instances	1,272,979	1,436,051	1,002,877	1,068,543
Meteorology	In-city sources	17	20	5	7
	In-city instances	116,867	106,614	30,305	55,632
	Nearby sources	177	195	115	122
	Nearby instances	1,006,814	1,108,873	626,418	665,463
Forecast	In-city sources	17	20	5	6
	In-city instances	390,702	361,624	106,380	51,870

The dataset has been used in papers [1][3] to infer the fine-grained air quality of current and future times. It can also be widely used in many urban computing scenarios introduced in paper [2] and machine learning tasks, such as multi-view learning, multi-task learning and transfer learning. Please cite the following three papers when using the dataset.

- [1] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, Tianrui Li. [Forecasting Fine-Grained Air Quality Based on Big Data](#). In the Proceeding of the 21th SIGKDD conference on Knowledge Discovery and Data Mining (KDD 2015).
- [2] Yu Zheng, Licia Capra, Ouri Wolfson, Hai Yang. [Urban Computing: concepts, methodologies, and applications](#). ACM Transaction on Intelligent Systems and Technology, 5(3), 2014.
- [3] Yu Zheng, Furui Liu, Hsun-Ping Hsieh. [U-Air: When Urban Air Quality Inference Meets Big Data](#). In the Proceeding of the 19th SIGKDD conference on Knowledge Discovery and Data Mining (KDD 2013).

1. City Data

Description:

This part of data is stored in “city.csv”, denoting the information of a city. As shown in Figure 1, there are 43 cities from two city clusters, where ‘Cluster A’ consists of 19 cities near Beijing and ‘Cluster B’ covers 24 cities near Guangzhou. Each pin point denotes a city.

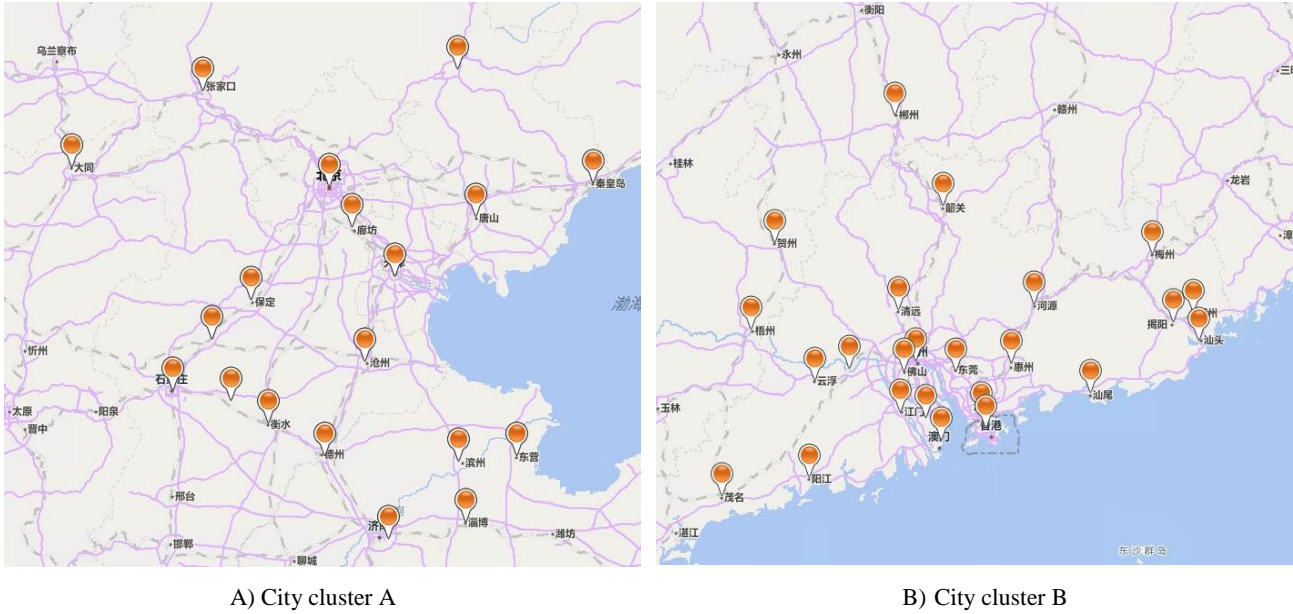


Figure 1 Geographical Distribution of 43 cities

Schema:

Each row in city.csv stands for a city, and the columns, separated by comma, are defined as follows:

City ID	Chinese Name	English Name	Latitude	Longitude	Cluster ID
---------	--------------	--------------	----------	-----------	------------

City ID is represented by 3 numbers.

English name is the Chinese Pinyin corresponding to the Chinese Name.

Latitude and longitude is the coordinate of the center of this city. We choose the town hall of a city as its center.

Cluster ID has two values, 1 means ‘City Cluster A’ and 2 means ‘City Cluster B’.

Example:

001, 北京, BeiJing, 39.904210, 116.407394, 1

004, 深圳, ShenZhen, 22.543099, 114.057868, 2

006, 天津, TianJin, 39.084158, 117.200982, 1

009, 广州, GuangZhou, 23.129110, 113.264385, 2

2. District Data

Description:

This part of data is stored in “district.csv”, denoting the information about 380 districts in 43 cities. Totally, there are 224 districts in ‘City Cluster A’ and 156 districts in ‘City Cluster B’. Most meteorological and weather forecast data are provided in a district level.

Schema:

Each row in district.csv stands for a district, and the columns, separated by comma, are defined as follows:

District ID	Chinese Name	English Name	City ID
-------------	--------------	--------------	---------

District ID is represented by 5 numbers and the first 3 numbers is same to the City ID.

English name is the Chinese Pinyin corresponding to the Chinese Name.

City ID means the city which this district belong to.

Example:

00101, 海淀区, HaiDianQu, 001

00102, 石景山区, ShiJingShanQu, 001

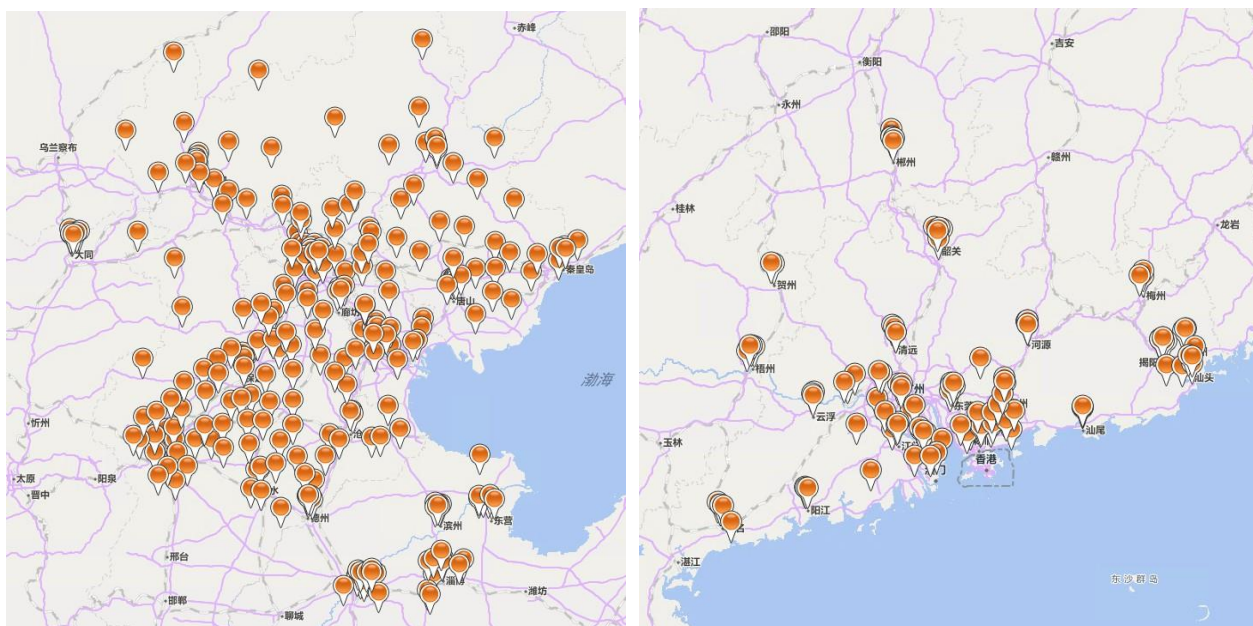
00103, 丰台区, FengTaiQu, 001

00104, 房山区, FangShanQu, 001

3. Air Quality Monitoring Station Data

Description:

This part of data is stored in “station.csv”, denoting the information of air quality monitoring stations in 43 cities. Totally, there are 437 air quality monitoring stations, where 271 stations are located in ‘City Cluster A’ and 166 stations in ‘City Cluster B’. Figure 2 plots all stations’ geographical distribution. Each pin point denotes an air quality monitoring station.



A) Stations in City Cluster A

B) Stations in City Cluster B

Figure 2 Geographical Distribution of 437 air quality monitoring stations

Schema:

Each row in station.csv stands for an air quality monitoring station, and the columns, separated by comma, are defined as follows:

Station ID	Chinese Name	English Name	Latitude	Longitude	District ID
------------	--------------	--------------	----------	-----------	-------------

Station ID is represented by 6 numbers, and the first 3 numbers is same to the City ID.

English name is the Chinese Pinyin corresponding to the Chinese Name.

District ID means the distinct that this station belongs to. It can be used to find the related meteorology and weather forecast data.

Example:

001001, 海淀北部新区, HaiDianBeiBuXinQu, 40.090679, 116.173553, 00101
001002, 海淀北京植物园, HaiDianBeiJingZhiWuYuan, 40.003950, 116.205310, 00101
001003, 石景山古城, ShiJingShanGuCheng, 39.914409, 116.184239, 00102
001004, 丰台云岗, FengTaiYunGang, 39.815128, 116.171150, 00103

4. Air quality Data

Description:

This part of data is stored in “airquality.csv”, denoting the air quality information of 43 cities. In total, there are 2,891,393 air quality records from 437 air quality monitoring stations over a period of one year.

Schema:

Each row in airquality.csv stands for an air quality record, and the columns, separated by comma, are defined as follows:

Station ID	Time	PM25	PM10	NO2	CO	O3	SO2
------------	------	------	------	-----	----	----	-----

Each air quality instance has six pollutants: PM25, PM10, NO2, CO, O3 and SO2, all these values are the concentration ($\mu\text{g}/\text{m}^3$) except of CO's unit is mg/m^3 . AQI (air quality index) can be calculated based on [HJ633-2012](#).

Example:

001001, 2014-05-01 00:00:00, 138, 159.4, 56.3, 0.9, 50.8, 17.2
001001, 2014-05-01 01:00:00, 124, 163.9, 38.7, 0.9, 51.1, 17.9
001001, 2014-05-01 02:00:00, 127, 148.4, 55.6, 1, 27.2, 16.6
001001, 2014-05-01 03:00:00, 129, 145.6, 65.7, 1, 9.7, 16.7

Statistics:

Take Beijing, Tianjin, Guangzhou and Shenzhen for example, the distribution of air quality in six AQI level is presented in Figure 3. Beijing and Tianjin have a worse air quality than Guangzhou and Shenzhen. Figure 4 plots the monthly average PM2.5 concentration From May, 2014 to April, 2015. PM2.5 concentration is higher when weather is cold.

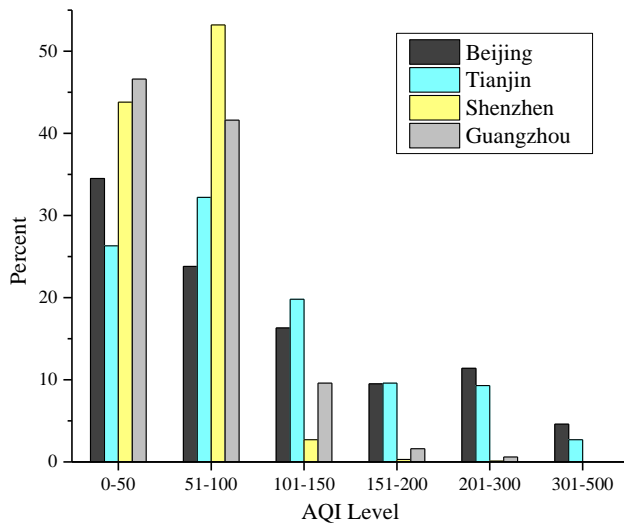


Figure 3 Distribution of air quality in 6 levels

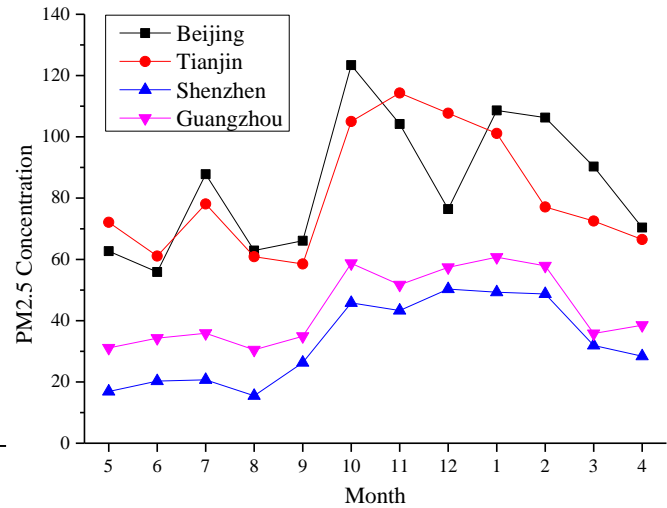


Figure 4 Monthly average air quality

Table 2 shows the percentage of missing values of 6 pollutants of Beijing. PM10 is missing seriously. It is caused by two reasons: one is the crash or error of crawlers; the other is incorrect data released by official data provider. This also bring some dirty data like outliers or duplicates after simple data cleaning in air quality data , meteorology data and weather forecast data. Missing value is represented by NULL in the data files.

Table 2 Percentage of Missing values of 6 pollutants in Beijing

Missing Value	PM25	PM10	NO2	CO	O3	SO2
Percent	13.3%	45.1%	16%	15.1%	15.4%	15.2%

5. Meteorological Data

Description:

This dataset is stored in “meteorology.csv”, denoting the real-time meteorology data in 43 cities. Totally, there are 1,898,453 meteorology records. Commonly, there are some district-level meteorology records and one city meteorology record in one city at one specific hour.

Schema:

Each row in meteorology.csv stands for a meteorology record, and the columns, separated by comma, are defined as follows:

ID	Time	Weather	Temperature	Pressure	Humidity	Wind Speed	Wind Direction
----	------	---------	-------------	----------	----------	------------	----------------

ID is same to the district id or city id.

Weather has 17 different values defined as follows:

0	1	2	3	4	5	6	7	8
Sunny	Cloudy	Overcast	Rainy	Sprinkle	Moderate rain	Heaver rain	Rain storm	Thunder storm
9	10	11	12	13	14	15	16	
Freezing rain	Snowy	Light snow	Moderate snow	Heavy snow	Foggy	Sand storm	Dusty	

Temperature is in the format of Celsius, unit is $^{\circ}\text{C}$.

Pressure means surface pressure, unit is hPa.

Humidity means relative humidity.

Wind speed's unit is m/s.

Wind direction has 10 different values defined as follows:

0	1	2	3	4	9	13	14	23	24
No	East	West	South	North	Unstable	Southeast	Northeast	Southwest	Northwest

Example:

001,2015-04-30 22:00:00,5,24.4,1006,50,3.5,13

001,2015-04-30 23:00:00,1,20.5,1007,67,1.9,23

00101,2014-05-01 00:00:00,0,20,1004,56,7.92,13

00101,2014-05-01 01:00:00,0,18,1004,64,7.56,13

Statistics:

Figure 5 shows the distribution of different weather of Beijing in one year. There are almost 10% foggy and dusty data.

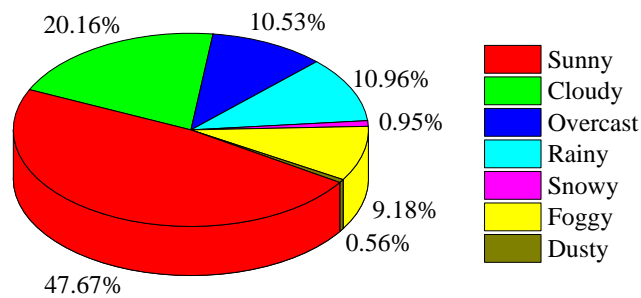


Figure 5 Distribution of weather

Table 3 shows the percentage of missing values of 6 weather conditions of Beijing.

Table 3 Percentage of Missing value of 6 weather conditions in Beijing

Missing Value	Weather	Temperature	Pressure	Humidity	Wind Speed	Wind Direction
Percent	24.2%	22.3%	40.1%	21.7%	29.6%	21.7%

6. *Weather forecast Data*

Description:

This dataset is stored in “weatherforecast.csv” which denotes the forecasted future weather in the following two days. The updating frequency of the forecasts has 3 different temporal granularity, some are updated every 3 hours, and some are 6 hours, 12 hours. Totally, there are 910,576 weather forecast records in 4 major Chinese cities. Like meteorology data, there are some district-level weather forecast records and one city weather forecast record in one city at one specific hour.

Schema:

Each row in the file stands for a weather forecast instance, and the columns, separated by comma, are defined as follows:

ID	Forecast Time	Future Time	Temporal Granularity	Weather	Up temperature	Bottom Temperature	Wind Level	Wind Direction
----	---------------	-------------	----------------------	---------	----------------	--------------------	------------	----------------

ID is same with the district id or city id.

Wind level has values like 3.5, 4.5, 5.5... We take the media value 3.5 to represent wind level as it always has two levels like 3-4 level in weather forecast.

Weather and wind direction are same to the same weather conditions in Meteorology.

Example:

```
001,2015-04-30 07:00:00,2015-04-30 08:00:00,3,1,28,21,3.5,3
001,2015-04-30 18:00:00,2015-04-30 20:00:00,3,14,25,22,3.5,3
00101,2014-05-01 00:00:00,2014-05-01 02:00:00,6,2,19,16,0,3
00101,2014-05-01 00:00:00,2014-05-01 08:00:00,6,2,29,19,0,3
```

Statistics:

Table 6 shows the distribution of temporal granularity of Beijing. Most weather forecast records have a temporal granularity of 3 and 12.

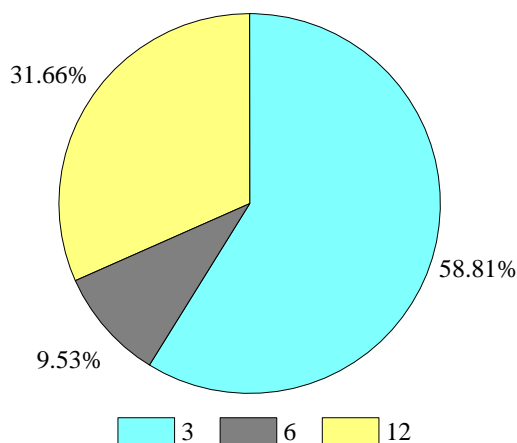


Figure 6 Distribution of temporal granularity

Contact: Yu Zheng, Lead Research at Microsoft Research

Email: yuzheng@microsoft.com

Homepage: <http://research.microsoft.com/en-us/people/yuzheng/>